Network Visualization of Car Inspection Data using Graph Layout

Jaakko Talonen, Miki Sirola
Aalto University, Department of Information and Computer Science
Espoo, Finland
jaakko.talonen@aalto.fi, miki.sirola@aalto.fi

Mika Sulkava
MTT Agrifood Research Finland, Economic Research
Helsinki, Finland
mika.sulkava@mtt.fi

***Abstract*—In this paper, we introduce the network visualization based on the rejection reasons on car inspections. It is compared with the visualization based on principal component analysis. The largest private provider of vehicle inspections in northern Europe, A-Katsastus, published rejection statistics in Finland for the fourth time. The statistics is published in dozens of tables on the basis of the year of introduction into use, make or model. Our goal is to visualize all this information in one network. The car inspection data is aggregated with the produced visualization. However, the dependencies between the different rejection reasons and cars can be efficiently studied by exploring our network visualization.**

Keywords-Car inspection; Rejection reason network; Visualization; Gephi; ForceAtlas2

I. INTRODUCTION

In our research, we used a desktop application Gephi. It is an interactive visualization and exploration platform [1]. It is commonly used for the visualization of social networks [2], Facebook friends, Twitter, etc. For example, the discussions between the Members of the Parliament in Finland have been visualized [3]. In that visualization, those who took part to the same discussions are connected. Visualizations of Social Networks have been widely studied, e.g., in [4], a network of social relations is created using publicly articulated mutual "friendship" links. In this paper, we present results suggesting that network visualization is suitable also for car inspection data where different rejection reasons are presented as "bridges" between the cars.

Typically, in reliability rating reports, areas of inspection are excluded and the most common grounds for rejection to keep legibility simple enough on the different tables, e.g., TÜV reports [5]. Only rank, car make – model and fault percentage in different "introduced into use" tables are shown. It means that there are 40 different tables published per year, i.e., 320 tables since 2004. Our goal is to aggregate all this data to one visualization including the most common grounds for rejection. It helps the users to make their own conclusions by extracting some rejection reasons, such as tires, which are totally dependent on drivers when analyzing the data. Our network visualization is compared with a visualization generated by Principal Component Analysis (PCA) which was used on our earlier work [6].

II. DATA AND PREPROCESSING

A-Katsastus Group inspected approximately 925 000 passenger cars in Finland in 2011 [7]. This and previously published data is used in our visualization. Yearly data is divided into several tables by the age of the car. A similar publication has previously been produced on the basis of their statistics by the Swedish company Bilprovningen and the German vehicle inspection chain Dekra.

The top three rejection reasons (RR) are listed, if a certain car is inspected more than $N=100$ times and the same rejection reason is listed more than $L=10$ times. The proportion of these requirements is $p=0.1$. In theory $p \in]0,1]$, but in practice it can be assumed that p varies around value 0.1. In this paper "car" means its model, type and age. In Finland, new cars are inspected on third and fifth year and older cars yearly. Newer cars have fewer rejections. Therefore there is less information about new cars than old ones. Also an average rejection r [percentage] is listed.

In the original data rejections are divided into 13 different classes, such as tires, brakes, steering and control devices; see Table 1. For analysis, a 14th class is defined as "unclassified" reason, because mostly the new cars have no listed rejection reasons in the data.

TABLE I. REJECTION REASONS

Classified rejection reasons and the sum of most popular reasons for rejection [A-Katsastus]		
RR	***ID***	***# 1st RR (2011)***
chassis	1	9
front suspension	2	73
shock absorption	3	17
suspension	4	12
brakes	5	163
other equipment	6	0
steering and control devices	7	53
exhaust emissions	8	102
tires	9	0
parking brake	10	22
rear axle	11	18
airbags	12	0
identification number	13	0

 ISBN: 978-1-61208-242-4

Data is quantified for the analysis using the rejection percentage *r* and the rejection reasons. We define a car matrix as

$$X = \begin{bmatrix} x_{1,1} & \cdots & x_{1,m} & x_{1,m+1} \\ \vdots & \ddots & & \\ x_{n,1} & & x_{n,m} & x_{n,m+1} \end{bmatrix}, \quad (1)$$

where n is the number of different cars and m the number of classes. In quantification it is assumed that the k^{th} RR *j* has probability

$$x_{i,j} = (p + a(k)) \cdot r, \quad (2)$$

where vector *a* is defined in this research as:

- All three reasons are listed: a = [0.04 0.02 0],
- Two reasons are listed: a = [0.02 0],
- Only one reason is listed: a = 0.

So, if all three rejection reasons are listed for the car *i*, the row sum of the first 13 cells for this car is *(0.14+0.12+0.1)r = 0.36r* with the assumption *p=0.1* (based on the A-Katsastus publication requirements). Last cell *x(i,14) = 0.64r*, so the sum of each row X is *r*. If no reasons are listed for a car, then *x(i,14) = r*. In practice this means that we assumed that the most common (listed first) reason is *2r* percentage units more probable than the second listed reason.

In the year 2008, RRs were classified in a different way than in the years 2009 - 2011. Therefore, it was excluded from the analysis and three matrices X(2009), X(2010) and X(2011) were visualized. As mentioned before, newer cars are inspected every second year, so in matrices X(2009) and X(2011) there are cars, which are missing from the matrix X(2010). These data matrices were combined row by row. In a new matrix Z one row represents one car. Older statistics have less effect on the model and it is defined as

$$Z(i,:) = \frac{\sum_{k=2009}^{2011} \lambda^{2011-k} L(i,2011-k) X_k(i,:)}{L(i,1) + L(i,2)\lambda + L(i,3)\lambda^2}, \quad (3)$$

where $\lambda=[0,1]$ is a forgetting factor. If $\lambda=\alpha$, it is expected that α% of last years car RRs are taken into account in the visualization. *L* is a zero-one vector defining if car data exists on matrix X(k) or not. Car *i* introduced in use, e.g., in 2007 was inspected in the year 2010, but not in the years 2009 and 2011, so then *L(i,:) = [0 1 0]*. Zero values in matrix Z mean that there is no connection between the car *i* and RR *j*.

III. METHODS

A. Principal Component Analysis

Principal Component Analysis (PCA) is a method for orthogonal linear transformation. The dimension of the data is reduced by transforming it to a new coordinate system such that the greatest variance lies on the first component [9]. The quantified matrices are combined as

$$C = \begin{bmatrix} X_{2009} & X_{2010} & X_{2011} \end{bmatrix}^T, \quad (4)$$

and matrix C is projected to subspace by placing the first N principal components in matrix

$$\Theta = (\theta_1 | \cdots | \theta_N). \quad (5)$$

B. Network Visualization

Our network layout is based on the ForceAtlas2 (FA2) method [8]. It is suitable for graphs with 10 to 10000 nodes. Cars and RRs are represented by colored balls in a graph. The attraction force *F* between two nodes *n(1)* and *n(2)* depends linearly on the distance *d(n(1),n(2))*.

FA2 is a continuous algorithm and the model is based on attraction and repulsion proportional to distance between nodes. Various layouts are achieved with different initial coordinates and parameter settings, see (2, 3). The main goal is to produce a readable spatialization and devise an energy model that could be easily understood by users. A clear visualization where nodes are separated and not overlapped (this feature can be forced in Gephi [1]) is reached by various parameter settings. In our model default values of scaling were increased and gravity decreased to obtain a sparser graph.

IV. EXPERIMENTS

A. Principal Component Analysis

We tested two different methods for car data visualization. Matrix *C* was projected to a 2-dimensional space using PCA. First and second components were visualized using Google Motion Chart where the RR statistics for 2009-2011 can be interactively explored. Cars with the same RR are situated in the same place in coordinates. However, this visualization is not very practical. Cars with only one listed RR are situated in corners, because *m* in (1) relatively small [6]. Readibility of the graph is not very good if only one plot is used, because RRs are shown only in a loadings plot, see Figure 1.

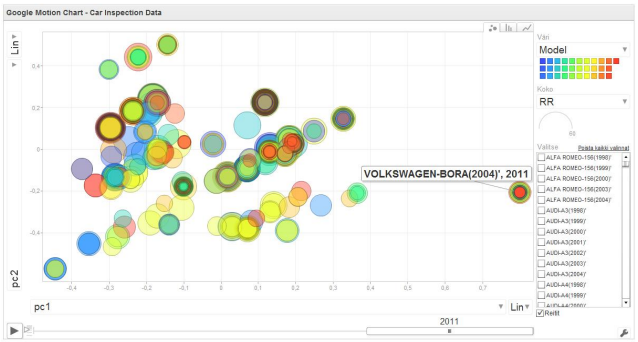

Figure 1. PCA results are shown in the motion chart. Volkswagen Bora (2004) in the last year's statistics (2011) is projected to the far right. The first component loading for exhaust problem is positive.

B. Network Visualization

All cars are connected to rejection results based on matrix *Z*, see (3). ForceAtlas2 algorithm is used to order initial car coordinates to stabile positions. Some of the

default parameters were changed: *Edge Weight Influence = 0.02, Scaling = 50, Gravity = 1.0 and Tolerance = 0.1*. Cars with same rejection reasons are situated in the same area in the graph.

Some cars have no listed rejection reasons. Therefore, the unclassified class *x(i,14)* was introduced in the previous section. Cars without RR information are located in the same area. The total weights of RRs are the same as the average rejection percentage *r*. Each car type and model is connected to the same cars with up to two years difference in age with small connection weights. With one year difference *w=0.2* and two years difference with *w=0.1*. By this procedure, the visualization is more informative, because a relatively large cluster of nodes with only one connection to unclassified RR is avoided. In practice, it means that in a graph we are assuming that if car *i* does not have listed RRs, it has with small probability the same RRs than older or newer cars with the same model and type, but are not listed in the published statistics.

In the past three years, there were 1060 cars which were inspected more than 100 times. It means that the size of matrix Z is *(n=1060, m=14)* in our visualizations. A produced network with *(λ = 0.8)* is visualized in Figure 2. All cars are connected to unclassified RR (ID=14) which is situated in the center of a graph. Infrequent RRs are placed on graph borders and RR(ID=i) which is dependent with another RR(ID=j) near each other.

The parking brake RR is mentioned in car inspection statistics in highlighted cars shown in Figure 3. Chassis, front suspension and brakes problems also occur rather probably in these cars, because these RRs are rather close each other in the graph.

Some of the RRs are totally driver dependent. Rather old Toyota Corollas have had small rejection percentage *r*. In addition, one of the reasons was bald tires. The connections of Toyota Corolla (2006) are shown in Figure 4. About *3.5%* of the cars were rejected in the car inspections. Based on matrix Z, connections show that in the past three years 3rd top-rated reason was tires. Also, additional connections show that older and newer cars have had the same kind of RRs.

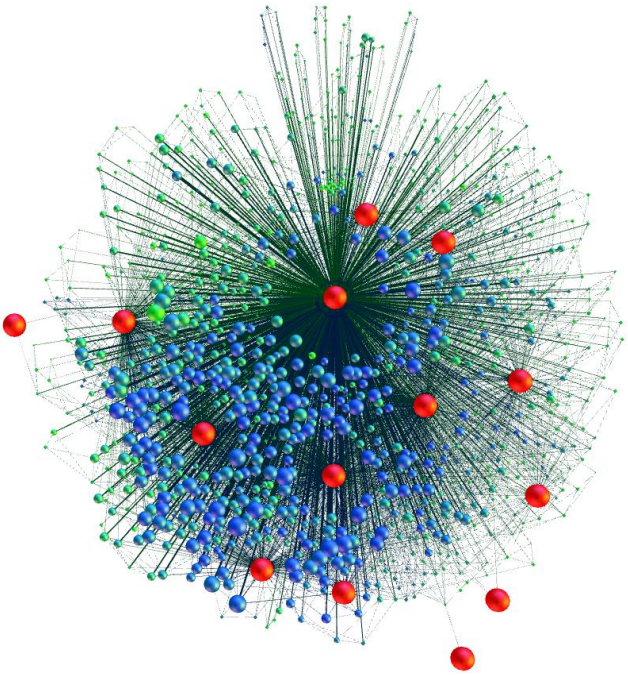

Figure 2. Cars and rejection reason classes (constant size red balls, m=14) are visualized. Old cars are represented by blue balls and newest (2008) by green balls, n=1060. Edges exist between RRs and cars. Also cars with age ±2 are connected. Sizes of the car balls and edges between cars and RRs are proportional with the rejection rates *r*.

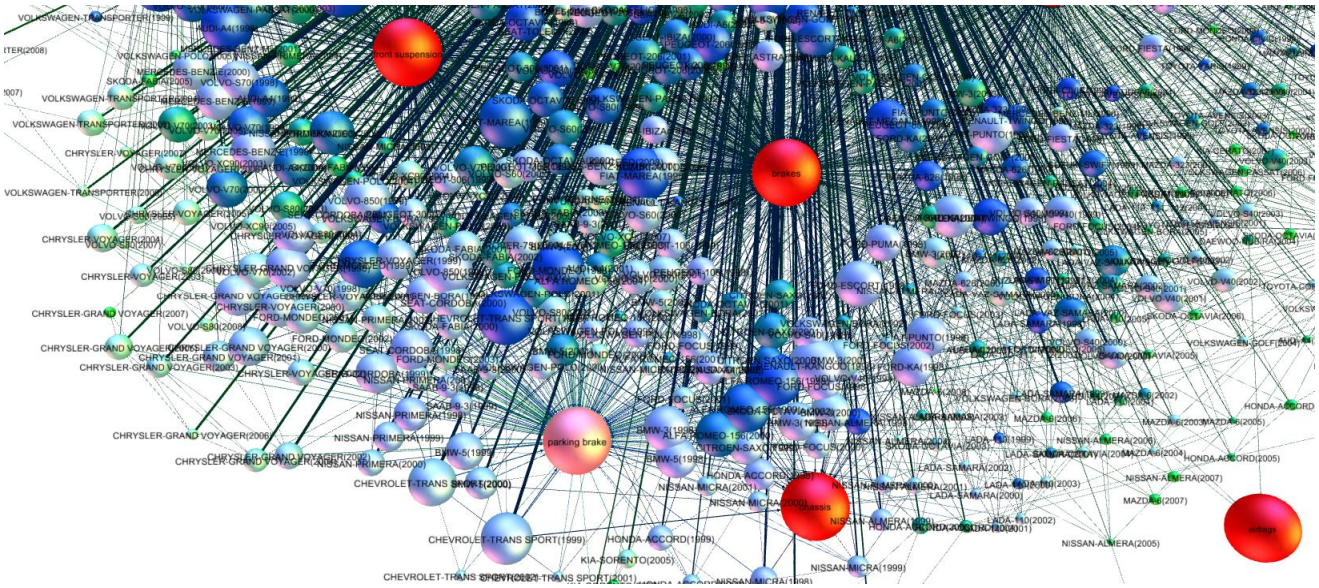

Figure 3. The labels of cars with the parking brake rejection reasons are highlighted.

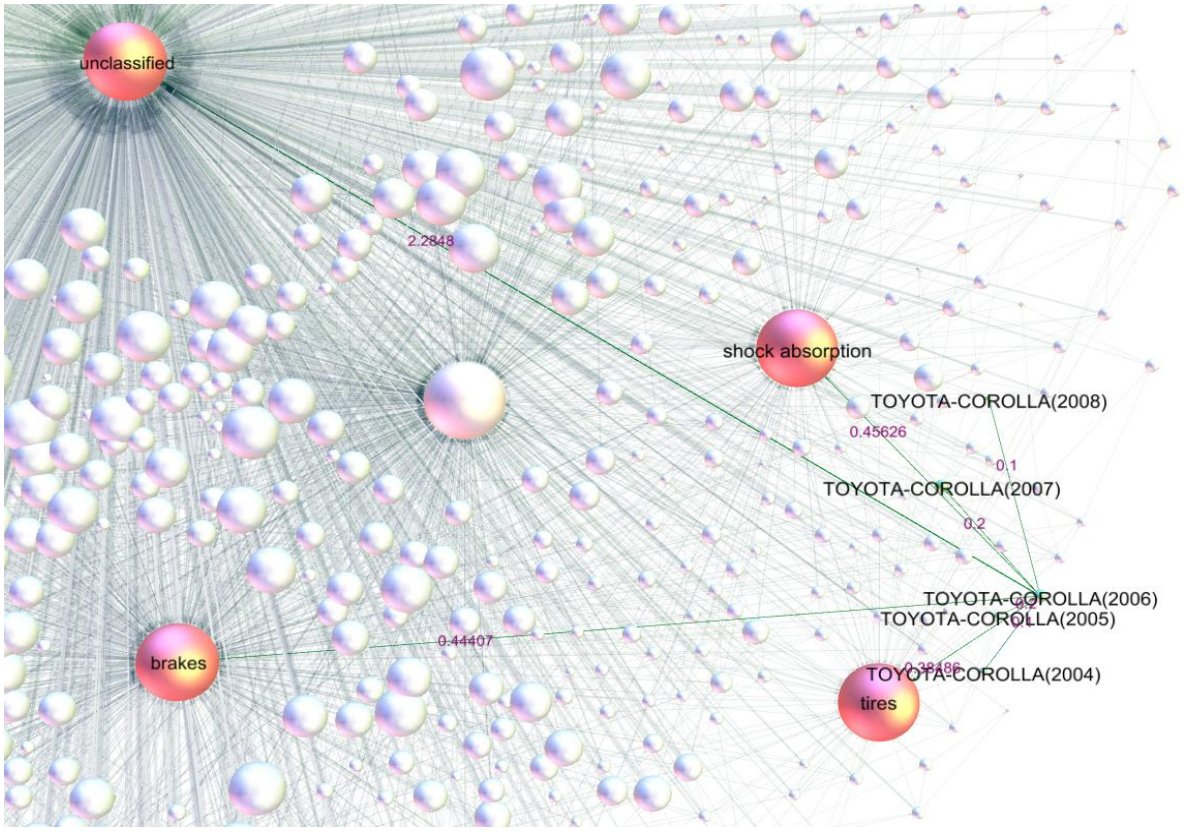

Figure 4. The connections of Toyota Corolla (2006) are shown. It is connected with all rejection reasons mentioned in reports 2009-2011 and the same cars which have been introduced into use between 2004 and 2008. Short edges between Toyota Corolla cars are corresponding similar RRs.

V. CONCLUSION AND FUTURE WORK

Our goal in information visualization was reached even though the provided data was not complete. The achieved results with certain assumptions related to data preprocessing and visualization are reported in this paper. PCA visualization was found to facilitate data exploration to some degree. Exploring the car inspection data is faster using the visualizations in Gephi or in a browser than by the dozens of tables. Visualization with the same parameters as presented in this paper is available on web [10]. A user can study the dependencies between the different rejection reasons and cars by exploring our network visualization.

Our work is still in progress and in future work, we will use additional car inspection data to get more reliable and high quality visualizations. Other layouts will be considered in order to improve the network readability, e.g. [11].

ACKNOWLEDGMENT

We would like to thank Aalto University, Helsinki Doctoral Programme in Computer Science - Advanced Computing and Intelligent Systems (Hecse) and A-Katsastus for providing the data.

REFERENCES

[1] M. Bastian, S. Heymann and M. Jacomy, “Gephi: an open source software for exploring and manipulating networks”, International AAAI Conference on Weblogs and Social Media, 2009.

[2] N.B. Ellison et al., “Social Network Sites: Definition, History, and Scholarship”, Journal of Computer-Mediated Communication, Wiley Online Library, vol. 13, pp. 210–230, 2007.

[3] Discussion Network of the Members in The Finnish Parliament, 2012 http://gexf-js.teelmo.info/index.html#edustajien-puheet-network-2012-party_v2_gt3.gexf, retrieved on June 2012.

[4] J. Heer and D. Boyd, D, “Vizster: Visualizing Online Social Networks”, IEEE Symposium on Inf. Visualization, pp. 32–39, 2005.

[5] TÜV reports http://www.anusedcar.com/, retrieved on June 2012.

[6] J. Talonen and M. Sulkava, “Analyzing Parliamentary Elections Based on Voting Advice Application Data”, Advances of Intelligent Data Analysis, Springer, pp. 340–351, 2011.

[7] http://www.a-katsastus.com/, retrieved on June 2012.

[8] M. Jacomy and T.Venturi, “ForceAtlas2, A Graph Layout Algorithm for Handy Network Visualization”, unpublished 2011.

[9] J. Hair, R. Anderson, R. Tatham and W. Black, “Multivariate Data Analysis” Prentice Hall, 5th edition, 1998.

[10] Exported car inspection data network visualization http://dl.dropbox.com/u/7846727/2012gephi/index.html#cars.gexf, retrieved on June 2012.

[11] A.J. Enright and C.A. Ouzounis, “BioLayout – an automatic graph layout algorithm for similarity visualization, Bioinformatics, vol. 17, pp. 853–854, 2001.

 ISBN: 978-1-61208-242-4